

Rumor Detection on Twitter Pertaining to the 2016 U.S. Presidential Election

Zhiwei Jin^{1,2}, Juan Cao¹, Han Guo^{1,2}, Yongdong Zhang¹, Yu Wang³ and Jiebo Luo³

¹Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, CAS, Beijing 100190, China
²University of Chinese Academy of Sciences, Beijing 100049, China

³University of Rochester, Rochester, NY 14627, USA
{jinzhiwei, caojuan, guohan, zhyd }@ict.ac.cn; ywang.tsinghua@gmail.com; jluo@cs.rochester.edu

Abstract

The 2016 U.S. presidential election has witnessed the major role of Twitter in the year's most important political event. Candidates used this social media platform extensively for online campaigns. Millions of voters expressed their views and voting preferences through following and tweeting. Meanwhile, social media has been filled with fake news and rumors, which could have had huge impacts on voters' decisions. In this paper, we present a thorough analysis of rumor tweets from the followers of two presidential candidates: Hillary Clinton and Donald Trump. To overcome the difficulty of labeling a large amount of tweets as training data, we first detect rumor tweets by matching them with verified rumor articles. To ensure a high accuracy, we conduct a comparative study of five rumor detection methods. Based on the most effective method which has a rumor detection precision of 94.7%, we analyze over 8 million tweets collected from the followers of the two candidates. Our results provide answers to several primary concerns about rumors in this election, including: which side of the followers posted the most rumors, who posted these rumors, what rumors they posted, and when they posted these rumors. The insights of this paper can help us understand the online rumor behaviors in American politics.

Introduction

Online social media services, such as Twitter, play an important role in various political events in recent years. Especially in the 2016 U.S. presidential election, Twitter became a primary battle ground: candidates and their supporters were actively involved to do campaigns and express their opinions by tweeting. For example, the Twitter account of Hillary Clinton has 12 million followers and her most popular tweet received more than 600,000 retweets and one million likes.

Meanwhile, the fact that various fake news and rumors were spreading on social media during the election became a serious concern. Many news articles criticized that fake news on social media may have influenced the election. From the popular rumor debunking website Snopes.com, we

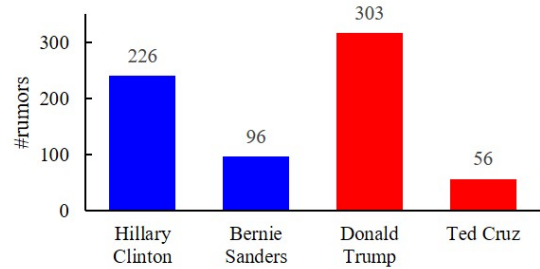


Figure 1: The number of rumors for popular presidential candidates on Snopes.com.

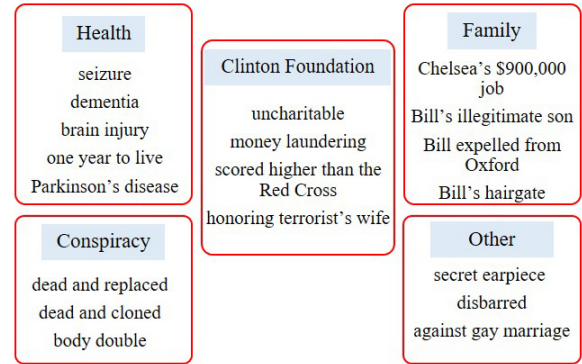


Figure 2: Various types of rumors related to Hillary Clinton.

find that many rumors coming from social media are related to the presidential candidates. Figure 1 illustrates the number of rumors about four popular candidates. Among all the 1,723 checked rumors from Snopes.com, 303 rumors are about Donald Trump and 226 rumors are about his Democratic opponent Hillary Clinton. There are 681 rumors related to these four candidates in total, which constitute 40% of all checked rumors. In Figure 2, we list the rumors about Hillary Clinton in categories. These rumors cover various topics to question her competence as a president. Undoubtedly, spreading them would have negative impacts on her candidacy.

In this paper, we aim to understand the rumor spreading

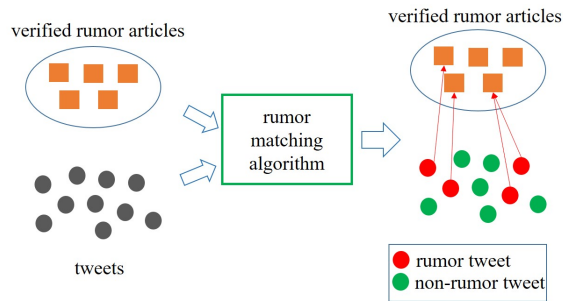


Figure 3: Rumor detection as a text matching task.

behaviors of candidates’ followers. A rumor is defined as a controversial and fact-checkable statement (DiFonzo and Bordia 2007). We use the checked rumors from Snopes.com as the objective rumor source in this paper. Before any further analysis, we need a method to detect rumors effectively and precisely on Twitter. Existing machine learning methods for rumor detection (Castillo, Mendoza, and Poblete 2011; Wu, Yang, and Zhu 2015; Jin et al. 2016a) fail to meet the requirement of this task for several reasons: 1) These methods need extensive labeled training data, which is expensive to label for the rumor detection problem; 2) Their computation efficiency is low especially when dealing with thousands of millions of tweets; 3) It is difficult to tell what types of rumors are posted as their binary results are not easily interpretable.

Considering these limitations, we propose to detect rumors as a text matching task in this paper (Figure 3). In this scheme, a set of verified rumor articles are collected as gold standard samples for reference. Each tweet is compared with these verified rumors to see if they match closely. As text matching is a well-studied problem in information retrieval, many unsupervised algorithms, such as tf-idf word vector model and word embedding, can be efficiently employed. Moreover, the matching result can not only reveal whether a tweet is rumor but also tell us which rumor the tweet is related to. This approach is better than the existing machine learning approaches in the aspects of efficiency and interpretability.

We use all the 1,723 checked rumors from Snopes.com as the verified rumor articles. In order to find the best matching algorithm, we conduct a comparative study of several competing algorithms. These algorithms are executed on a reasonably sized set of 5,000 manually labeled tweets to provide a fair performance comparison.

We then detect rumors with the selected most effective matching algorithm on over 8 million of tweets from 14,000 followers of the two primary presidential candidates: Hillary Clinton and Donald Trump. We inspect the rumor detection results from different aspects to answer following questions: which side posted the most rumors? who posted these rumors? what rumors did they posted? when did they post rumors? These insights help us understand the rumor tweeting behaviors of different groups of followers and can be helpful for mining voters’ real intentions and accurately detecting

rumors during political events in the future.

In summary, our work makes three main contributions:

1. We formulate the rumor detection problem on Twitter as an unsupervised matching task. We use correctly verified rumors on Snopes.com as gold standard samples for rumor matching. Compared with other existing machine learning approaches, this scheme is more efficient and interpretable.
2. We conduct an empirical study of several competing matching algorithms. After comparing the performance of five different unsupervised matching algorithms on a reasonably sized labeled set, we select the BM-25 method for rumor tweet detection, which can reach a precision of 94.7%.
3. We give a thorough analysis on rumor tweets detected on 8 million tweets collected from candidates’ followers. We produce insights into followers’ rumor spreading behaviors from various aspects.

Related Work

Online social media have gained huge popularity around the world and become a vital platform for politics. With thousands of millions of tweets available, many existing studies focus on sentiment analysis (Wang et al. 2012) and election prediction (Tumasjan et al. 2010) on Twitter. However, the openness and convenience of social media also fosters a large amount of fake news and rumors which can spread wildly (Friggeri et al. 2014). Compared with existing rumor detection works that are focused on general social events or emergency events (Jin et al. 2014), this paper presents a first analysis of rumors in a political election.

Most existing rumor detection algorithms follow the traditional supervised machine learning scheme. Features from text content (Castillo, Mendoza, and Poblete 2011; Kwon et al. 2013), users (Morris et al. 2012), propagation patterns (Wu, Yang, and Zhu 2015) and multimedia content (Jin et al. 2015; Jin et al. 2016b) are extracted to train a classifier on labeled training data. Some recent works further improve the classification result with graph-based optimization methods (Gupta, Zhao, and Han 2012; Jin et al. 2014; Jin et al. 2016a). Although machine learning approaches are very effective under some circumstances, they also have drawbacks. The supervised learning process requires a large amount of labeled training data which are expensive to obtain for the rumor detection problem. They are often computationally expensive, especially when dealing with thousands of millions tweets. They derive features in a “black box” and the classification results are difficult to interpret.

To overcome the time efficiency and interpretability problems of supervised learning, Zhao, Resnick, and Mei (2015) proposed a lexicon-based method for detecting rumors in a huge tweet stream. They extracted some words and phrases, like “rumor”, “is it true”, “unconfirmed”, for matching rumor tweets. Their lexicon is relatively small, thus the detection results tend to have high precision but low recall of rumors.

In this paper, we formulate the rumor detection as a text matching task. Several state-of-the-art matching algorithms

are compared for rumor detection. TF-IDF (Salton, Fox, and Wu 1983) is the most commonly used method for computing documents similarity. BM25 algorithm (Robertson and Zaragoza 2009) can be seen as an improved version of TF-IDF. Both are document relevance ranking models based on term matching. Recent research in deep learning for text representation embeds words or documents into a common vector space. Word2Vec (Mikolov et al. 2013) and Doc2Vec (Le and Mikolov 2014) are two widely used embedding models at the word and paragraph levels, respectively. In this paper, we give a comparative study of these methods to find the most suitable model for our rumor detection task.

Dataset

We collect a large-scale dataset for analyzing rumors during the 2016 U.S. presidential election from Twitter. For reliable rumor detection, we obtain a set of verified rumor articles from Snopes.com. We also manually construct a testing set to fairly evaluate the rumor detection methods.

The amount of tweets on Twitter is enormous and growing every second. It is almost impossible and also unnecessary to process all the data. As we are interested in finding out rumor spreading behaviors of different voter groups during the 2016 U.S. presidential election, we design a data sampling scheme to collect the tweets from the followers of the two primary candidates.

Using the Twitter API, we collect all the users who are following the Democratic presidential candidate Hillary Clinton and the Republic presidential candidate Donald Trump. We randomly select about 10,000 followers from each candidate’s follower list, which contains millions of followers. We remove users who follows both candidates to make sure each user follows only one specific candidate in our dataset. Users whose tweets are not available due to their privacy setting are also excluded. We then collect up to 3,000 most recent tweets for each user using the Twitter API. Altogether, we have over 8 billion tweets from 14,000 followers of the two candidates in our dataset (Table 1).

We collect a set of verified rumor articles from Snopes.com as gold standard samples for rumor matching. Snopes.com is a very popular rumor debunking website. Social media users can nominate any potential rumor to this site. The employed analysts then select some of these controversial statements to fact-check them as rumors or truth. An article is presented for each checked rumor by these professional analysts, which gives conclusion of the rumor followed by full description, source, origin, supporting/opposing evidences of the rumor story. We collect the articles of all the 1,723 checked rumors on this website to form the verified rumor article set.

To quantitatively evaluate the performance of rumor detection methods, we build a manually labeled tweet set. We randomly select 100 rumors from the verified rumor set. For each verified rumor article, we search our large tweet set with keywords extracted from the article. Each tweet in the search result is manually examined to check if it matches the rumor article. After these procedures, we obtain a set of 2,500 rumor tweets from 86 rumor articles. We then randomly samples the same number of unrelated tweets

Table 1: Dataset used for rumor analysis during election.

	#followers	#tweets
Hillary Clinton	7,283	4,452,087
Donald Trump	7,339	4,279,050
All	14,622	8,731,137

as negative samples. In this set, not only is each tweet labeled as rumor or not, but the rumor tweets are also labeled with their corresponding verified rumor articles. Therefore, we can perform both general rumor classification and fine-grained rumor identification with this dataset. The following is an example of a verified rumor article and three associated tweets.

Verified rumor article¹:

Shaky Diagnosis. A montage of photos and video clips of Democratic presidential candidate Hillary Clinton purportedly demonstrates she has symptoms of Parkinson’s disease. Photos and video clips narrated by a medical doctor demonstrate that Democratic presidential candidate Hillary Clinton likely has Parkinson’s disease. Hillary Clinton’s health has been the subject of intense speculation during the 2016 presidential campaign.....

Associated rumor tweets:

1. *Hillary collapse at ground zero! game over, Clinton! Parkinson’s blackout!*
2. *Wikileaks E-mails: Hillary looked into Parkinson’s drug after suffering from “decision fatigue”.*
3. *Exclusive Report: How true is this?? Hillary Clinton has Parkinson’s disease, doctor confirms.*

Rumor Detection

We formulate rumor detection on Twitter as a matching task in this paper (Figure 3). Compared with detection methods based on supervised machine learning, our scheme is more time efficient and interpretable, which is suitable for analyzing rumors from millions of tweets. With reliable rumor articles collected from Snopes.com, the key part of this scheme is the matching algorithm. In this section, we present a performance evaluation of several matching algorithms to find the best method for rumor detection task on Twitter.

Compared with the traditional rumor classification algorithms, our rumor matching scheme not only outputs a tweet as rumor or not but also identifies which rumor article it refers to if it is a rumor tweet. We perform comparative studies of different matching algorithms on both the classification and the identification task of rumor detection.

Rumor Detection Algorithms

We compare the performance of five matching algorithms with respect to the rumor classification task. The first set of methods includes two widely used term-based matching methods: TF-IDF and BM25. The second set includes two recent semantic embedding algorithms: Word2Vec and

¹The full article is available at: <http://www.snopes.com/hillary-clinton-has-parkinsons-disease/>

Doc2Vec. The third set is a lexicon-based algorithm for rumor detection on Twitter stream.

TF-IDF (Salton, Fox, and Wu 1983) is a widely used model in text matching. In this model, both the tweets and the verified rumor articles are represented as a v -dimensional vector, where v is the size of the dictionary of the corpus. Each element in the vector stands for the TF-IDF score of the corresponding word in the text. TF is the term frequency. IDF score is the inverse document frequency, which is calculated on the whole corpus.

BM25 (Robertson and Zaragoza 2009) is also a text similarity computing algorithm based on the bag-of-words language model. It is an improvement of the basic TF-IDF model by normalizing on term frequency and document length. Both TF-IDF and BM25 have been widely used in many related studies.

Word2Vec (Mikolov et al. 2013) represents each word in a corpus with a real-valued vector in a common semantic vector space. Compared with traditional lexical-based matching models, this algorithm evaluates the quality of word representations based on their semantic analogies. The vector representation of Word2Vec is successfully used in many text mining applications. We use the pre-trained Word2Vec model on a corpus of 27 billion tweets. The word dimension is 200. To aggregate a presentation for a whole text, we take the average of word vectors in the text.

Doc2Vec (Le and Mikolov 2014) is also an embedding algorithm on the semantic space, which can directly learn the distributed representations of documents. We use all the tweets and rumor articles for the unsupervised training of the model after standard pre-processing. We use the default parameter settings as in (Le and Mikolov 2014). After training, tweets and verified rumors are represented as 400-dimensional vectors.

For Word2Vec and Doc2Vec, the matching score between a tweet and a rumor article is computed based on the cosine distance of their vector representations.

Lexicon matching (Zhao, Resnick, and Mei 2015) is a lexicon-based rumor detection algorithms for efficiently detecting in huge tweet streams. It mines a couple of signal words or phrases for recognizing prominent rumor tweets. We use the same set of regular expression patterns as in (Zhao, Resnick, and Mei 2015) to match rumor tweets.

Evaluation on Rumor Classification Task

TF-IDF, BM25, Word2Vec and Doc2Vec represents texts as numeric vectors. The similarity between a tweet and a verified rumor is computed as their matching score. By setting a threshold h for each method, we classify tweets with matching scores larger than h as rumor tweets. We can achieve different precision and recall of rumor classification by varying the threshold. We test all the four methods on the 5,000 labeled tweet set. Figure 4 illustrates the precision-recall curves of these four algorithms. The lexicon matching algorithm detects rumors by keywords matching, thus its result is actually fixed (as a single point in Figure 4).

The highlighted round points on each curve in Figure 4 are points where the F1-measures are maximized, at 0.758, 0.82, 0.764 and 0.745 for TF-IDF, BM25,

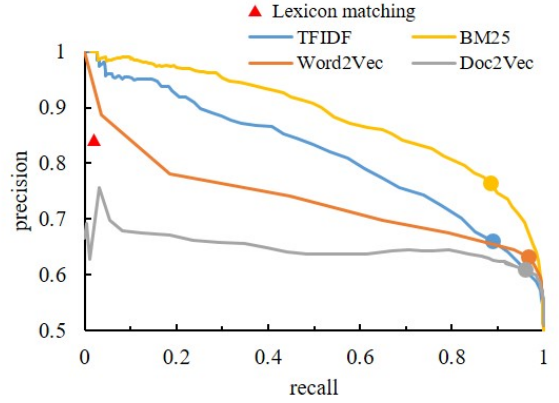


Figure 4: The comparative performance of four matching algorithms.

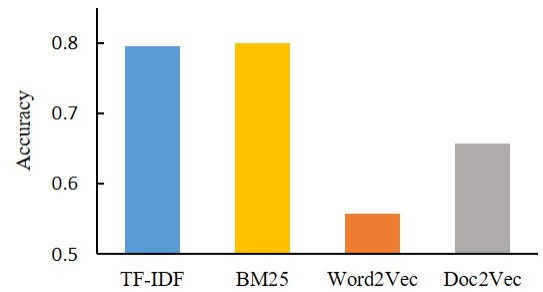


Figure 5: The accuracy of rumor identification task.

Word2Vec and Doc2Vec, respectively. The red triangle is the fixed result of lexicon matching. These results show that BM25 reaches the best performance among all the five rumor classification methods under different metrics. The two term-based methods (TF-IDF and BM25) outperforms the semantic-embedding and lexicon-based methods. For semantic-embedding, Word2Vec is slightly better than Doc2Vec. Lexicon matching can reach a rumor classification precision of 0.862, but its recall (0.008) is too low.

Evaluation on Rumor Identification Task

One extra advantage of our proposed rumor matching scheme is its ability to identify what rumor article a rumor tweet refers to, apart from classifying it as a rumor tweet. To compare the rumor identification performance of the four algorithms, we compute the similarity score between each pair of tweet and verified rumor article for the 2,500 labeled rumor tweets and 1,723 verified rumor articles. If the most similar rumor article of a tweet is exactly the same labeled rumor article for it, then this is an accurate rumor identification.

From the overall rumor identification accuracy of each rumor matching methods in Figure 5, we can observe that the BM25 algorithm achieves the best accuracy of 0.799. The accuracy of BM25 is only slightly better than that of TF-IDF,

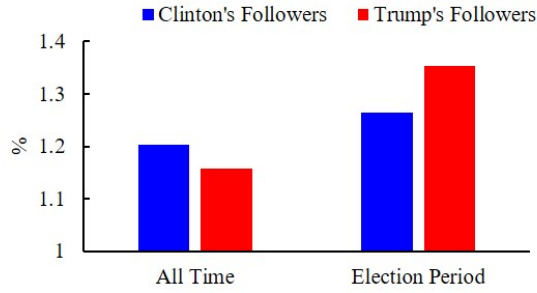


Figure 6: Rumor tweet ratio all time and in election period.

although it has major advantage in the rumor classification task. This is probably because BM25 can distinguish non-rumor tweets much better than TF-IDF. Another interesting finding is that Doc2Vec actually performs better on the rumor identification task than Word2Vec, although the latter has slightly better performance on the rumor classification task.

Rumor classification and rumor identification are correlated but also different tasks. With experiments on the labeled set, we find that the BM25 algorithm achieves the best performance on both tasks among others. For the following analysis on the whole 8 million tweet set, we use BM25 to detect rumor tweets and matching them with the corresponding rumor articles. To conduct a reliable and accurate analysis, we prefer a high precision for our rumor detection result. We set the similarity threshold $h = 30.5$ so that we can achieve a very high rumor classification precision of 94.7% and the recall of 31.5%.

Analyzing Rumor Tweets during the Election

This paper analyzes rumor tweets during the 2016 U.S. presidential election. For rumor analysis at a large scale, in this section, we use the BM25 rumor detection algorithm to detect rumor tweets from over 8 million tweets collected from the followers of Hillary Clinton and Donald Trump. Based on the results, we obtain insights into the rumor tweeting behaviors from various aspects.

Which side posted the most rumors?

Twitter has become an online battle field during the election. The number of rumor tweets reflects the involvement of candidates' followers in the election campaign. Which side of followers were involved most in spreading rumor tweets? To answer this question, we use the BM25 rumor classification method to detect rumors in the subset of tweets of the two candidates, respectively. Given our focus on rumors during the election period, we also analyze rumor tweets posted from April 2015² up to the present. Because the total number of tweets for each candidate's follower group is slightly different, we calculate the percentage of rumor tweets for a fair comparison.

²Clinton announced her candidacy on April 12, 2015 and Trump announced his on June 16, 2015.

The rumor tweet ratio all time and in the election period for Clinton and Trump are illustrated in Figure 6. We can observe from this figure that:

- Rumor tweets only constitute a small part on Twitter in general. Our rumor classification algorithm has a precision of 94.7% and a recall of 31.5% on the test set. Less than 1.2% tweets are detected as rumors in our 8 million tweet set. Based on these facts, we can assume that only about 4% tweets are rumor tweets on Twitter.
- All time, Clinton's followers are a slightly more active in posting rumor tweets than Trump's followers. 1.2% tweets are rumor tweets from Clinton's followers, which is about 4% more than that of Trump's followers.
- People tend to post more rumor tweets in the election time than in the whole time, especially for Trump's followers. Comparing their election period and all time rumor tweeting, Trump's followers have a rumor tweet ratio of 1.35% during the election, which is 18% higher than that in all time.
- During the election time, Trump's followers are more active in rumor tweeting than Hillary's followers. As the figure suggests, although Trump's followers are a slightly less active in rumor tweeting in the whole time, they become much more involved in posting rumors at the election time, compared with Clinton's followers.

Who posted these rumors?

Who are behind the rumors spreading on Twitter? We investigate this issue by analyzing rumor tweets posted by individual followers of the two candidates.

To obtain an overall impression of rumor spreaders, we rank users by the total number of rumor tweets they posted (Figure 7). From this analysis, we find that the number of rumor tweets posted by individual followers of each candidate follows a perfect power law distribution. This means that the majority of rumors are posted by only a few users on Twitter. To be exact, the top 10% users posted about 50% rumor tweets and the top 20% users posted about 70% of all rumor tweets. This is a common phenomenon revealed by both Trump's and Clinton's followers. We can conclude that rumor spreading on Twitter is mostly dominated by a small group of active users.

Moreover, Figure 7 also demonstrates the rumor tweeting pattern of each candidate's follower group from the user aspect. For example, we find that the Clinton's followers have a larger rumor tweet ratio than Trump's followers mostly because of the difference in their top-1000 followers (the magnified part in the figure).

Are these rumor-prolific followers just active in tweeting rumors or active in general tweeting as well? To understand this, we calculate the ratio of rumor tweets in all tweets posted by a user. In Figure 8, we show the rumor tweet ratio of the top 1000 users of Clinton's followers. We omit the figure of Trump's followers to save space, which have a very similar distribution.

In this figure, the x-axis is user ranked in the same order as in Figure 7 and the y-axis is the rumor tweet ratio of

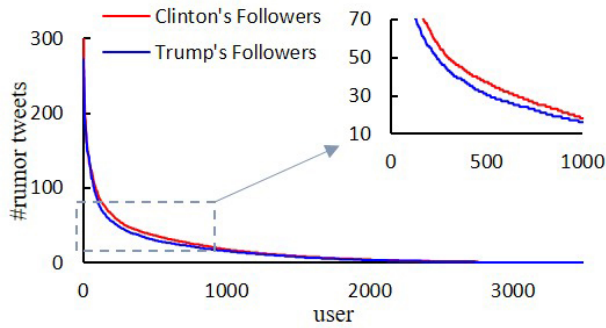


Figure 7: The number of rumor tweets posted by individual followers of Clinton and Trump. The plot on the top right is a magnification of the part in the dotted box.

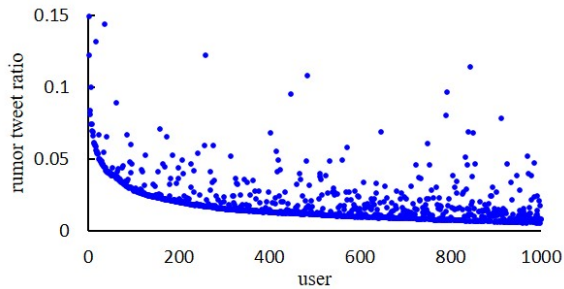


Figure 8: Rumor tweet ratio of Clinton's followers.

each user. We find that the rumor tweet ratio also follows a power law distribution except for some outliers. Followers who post more rumor tweets also tend to have a larger rumor tweet ratio. This means the rumor-prolific users did not randomly post any tweets; they were actually more concentrated on posting rumor tweets than the users who occasionally post a few rumor tweets. In fact, the rumor tweet ratio of the top 400 users is above 0.012, which is the average rumor ratio for Clinton's followers.

Case Study After analyzing rumor spreaders at a large scale, we can also conduct a detailed analysis for a specific user.

Take one of Trump's followers for example. This user posted 3,211 tweets in our dataset, 307 of which are detected as rumors. The rumor tweet ratio is as high as 9.6%, which means this user is very active in rumor tweeting. By examining the top keywords in all tweets posted by the user (Figure 9), we find this person is very focused on posting tweets about the 2016 presidential election: "Clinton", "Sanders", "Trump" and "election" are the most mentioned words in the tweets. After rumor detection, we find that the rumor tweets of this user are mainly about Clinton and Sanders rather than Trump: 15% tweets about Clinton and 28% tweets about Sanders are rumor tweets, while only 10% tweets about Trump are rumors. Given these observations, this user is probably a solid supporter of Trump and the Republic party during this election.

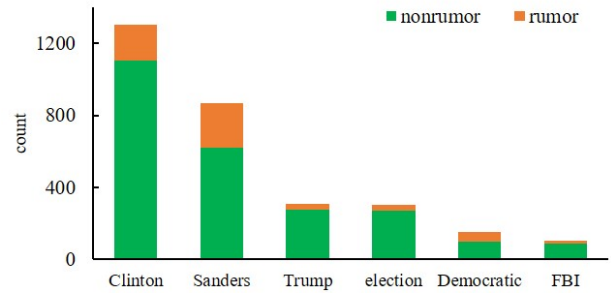


Figure 9: Keywords in the tweets posted by one of Trump's follower.

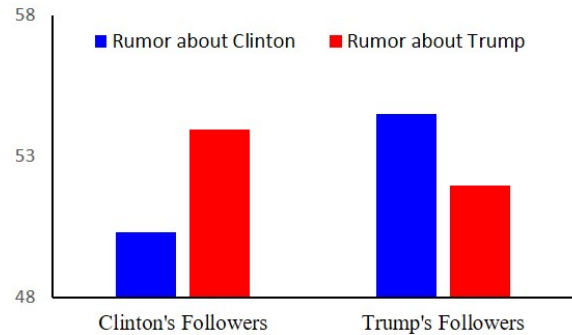


Figure 10: Rumors posted by followers of Trump and Clinton.

What rumors did they post?

Rumors cover various topics of interests. During the election, most rumors are focused on the candidates (Figure 1). People post different rumors to convey certain sentiments. By analyzing what people from different groups tweeted about in rumors, we can understand their intentions in this election.

We use BM25 to identify the content of each rumor tweet by matching it with the verified rumor articles from Snopes.com. As shown in Figure 5, this method has a rumor identification accuracy of 80%. Given our focus on the two primary presidential candidates, Hillary Clinton and Donald Trump, we only analyze rumor tweets related to them. Identified rumor tweets are grouped into two subsets based on which candidate their corresponding rumor articles refers to. After normalizing the number of candidate-related rumor tweets with the total number of rumor articles for this candidate in our dataset, we plot the rumor content spread by Trump's and Clinton's followers in Figure 10. We offer some analysis of this figure based on the normalized rumor tweet number.

First, both follower groups post rumors about their favored candidate as well as the opponent candidate. Rumors are often very controversial stories. They are normally negative towards one candidate during the election (Figure 2). Supporters of one candidate would spread rumors about the opponent as a negative campaign tactic and debunk rumors about their favored candidate. For example, we show two

Table 2: Top ranked rumor articles.

Top ranked rumors posted by Trump’s followers	
T1	Bill O’Reilly and Megyn Kelly are Clinton operatives.
T2	Hillary Clinton refused to congratulate American gold medalist Ginny Thrasher.
T3	Donald Trump: “laziness is a trait in blacks”
T4	Mother Teresa and Hillary Clinton’s talk on abortion.
T5	Donald Trump opposed the invasion of Iraq before it took place in 2003.
Top ranked rumors posted by Clinton’s followers	
C1	Donald Trump was nominated for a Nobel Peace Prize.
C2	Bill O’Reilly and Megyn Kelly are Clinton operatives.
C3	Hillary Clinton refused to congratulate American gold medalist Ginny Thrasher.
C4	Donald Trump denied saying that not paying federal taxes made him smart.
C5	Hillary Clinton didn’t publicly support same-sex marriage until 2013.

tweets about the rumor “Hillary Clinton has Parkinson’s disease” from our dataset:

Tweet 1: *Medical experts watching debate said Hillary showed “Telltale Signs” of Parkinson’s Disease.*

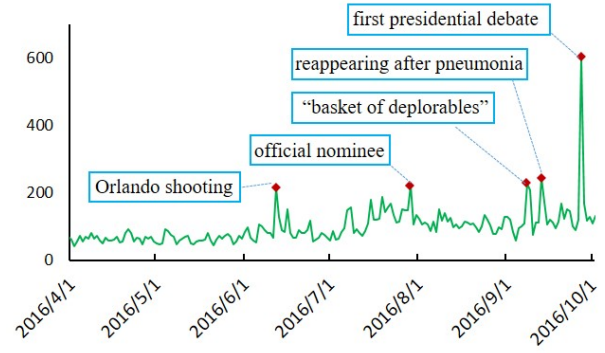
Tweet 2: *“I know her physician; I know some of her health history which is really not so good” Trump’s MD on Hillary—her MD shared her info with him?*

The first tweet comes from a follower of Trump. It is spreading the rumor by quoting medical experts. The second tweet comes from a follower of Clinton. It is obviously questioning the truthfulness of the rumor.

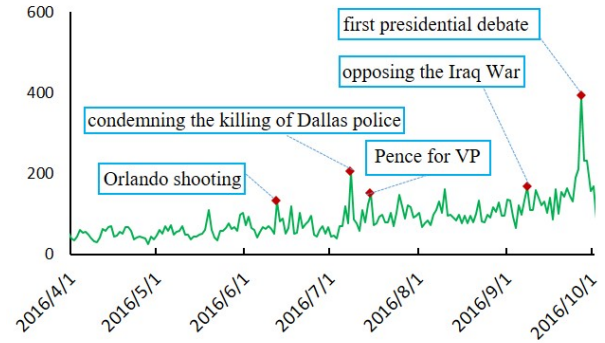
Second, users would post more rumor tweets about the opponent candidate than their favored candidate. Clinton’s followers post 8% more rumor tweets about Trump than rumors about Clinton. Trump’s followers post 5% more rumor tweets about Clinton than rumors about Trump. Moreover, Trump’s followers are more active in this rumor tweeting behavior towards both Clinton and Trump. The numbers of rumor tweets about the two candidates posted by Trump’s followers are both larger than those of Clinton’s followers.

Case Study To see what exactly is talked about in rumor tweets from the two follower groups, we rank rumor articles based on the number of matched tweets. The top ranked rumor articles about the two candidates are listed in Table 2. We can draw some conclusions from this table:

- These top ranked rumors are mostly about hot voting issues in this election: racial minorities (rumor T2, T3 and C3), mass media’s objectivity (T1 and C2), abortion (T4), LGBT (C5), military policy (T5) and candidate’s qualification (C1 and C5). Some rumors are so popular that they are ranked high in two follower groups (T1-C2 and T2-C3). Obviously, spreading rumors on these hot issues would influence people’s judgement on a candidate. Al-



(a) Rumor tweet timeline of Clinton’s followers



(b) Rumor tweet timeline of Trump’s followers

Figure 11: Rumor tweet timeline.

though some issues are more important in the election, such as economy and terrorism, few rumors are posted on them.

- Basically, Trump’s followers tweet more on rumors about Clinton and likewise. Apart from some shared rumors, the two follower groups have very different focuses on rumors. Taking the abortion rumor T4 as an example, there are 210 tweets about it from Trump’s followers and only 60 tweets from Clinton followers.

When did they post these rumors?

Timing is everything in politics. Analyzing the time patterns of rumor tweeting can reveal insights of online campaign. After counting the number of rumor tweets posted in each day, we plot the rumor tweeting of Clinton’s and Trumps followers over six months (April 2016 to September 2016). As Figure 11 shows: 1) Rumors in the earlier three months were relative smooth. As the election day approached, rumor tweeting became more and more intense; 2) The two follower groups had different time patterns of rumor tweeting: the fluctuation and peaks were different.

We annotate the key events for some rumor peaks in the figure to understand the inherent reason behind them. From September to April, the five marked events in Clinton’s timeline are: 1) the first presidential debate, 2) Clinton reappeared after diagnosed with pneumonia, 3) Clinton called

the Trump's supporters as "basket of deplorables", 4) Clinton was made the official party nominee, 5) the Orlando shooting. And the five marked peaks in Trump's timeline correspond to: 1) the first presidential debate, 2) Trump declared he was opposing the Iraq war before it started, 3) Mike Pence was announced as the vice president candidate, 4) Trump condemned the killing of five Dallas police officers, 5) the Orlando shooting.

We find that rumors peaked in three types of occasions: 1) key point in the presidential campaign, such as "the presidential debate" and "official nominee"; 2) controversial emergency events, including "the Orlando shooting" and "the Dallas shooting"; 3) events triggering rumors, such as "reappearing after pneumonia" and "opposing Iraq war". This insight reminds us to pay more attention to rumors during these types of events in the future political campaigns.

Conclusions

This paper studies the rumors spreading phenomenon on Twitter during the 2016 U.S presidential election. We propose a reliable and interpretable approach to detecting rumor tweets by matching them with verified rumor articles. After a comparative study, we find the BM25 matching algorithm outperforms other four competing rumor matching methods in this task. With a rumor detection precision of 94.7%, we use this method to detect rumors in over eight million tweets collected from the followers of the two primary presidential candidates. We provide a thorough analysis on the detected rumor tweets from the aspects of people, content and time. Many interesting rumor tweeting patterns are discovered, including: 1) more rumors are posted at election time than on the average; 2) rumor tweeting is dominated by a small group of users; 3) users post rumor related tweets to debunk rumors about their candidate or slander the opponent; 4) rumor tweeting erupts mainly in three types of occasions (key points in the presidential campaign, upon controversial emergency events, upon unique events), etc. These insights would help us better understand rumors during political events and inspire us to build more effective rumor detection algorithms in the future.

References

- [Castillo, Mendoza, and Poblete 2011] Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World Wide Web (WWW)*, 675–684. ACM.
- [DiFonzo and Bordia 2007] DiFonzo, N., and Bordia, P. 2007. *Rumor psychology: Social and organizational approaches*. American Psychological Association.
- [Friggeri et al. 2014] Friggeri, A.; Adamic, L. A.; Eckles, D.; and Cheng, J. 2014. Rumor cascades. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*.
- [Gupta, Zhao, and Han 2012] Gupta, M.; Zhao, P.; and Han, J. 2012. Evaluating event credibility on twitter. In *Proceedings of the SIAM International Conference on Data Mining*, 153. Society for Industrial and Applied Mathematics.
- [Jin et al. 2014] Jin, Z.; Cao, J.; Jiang, Y.-G.; and Zhang, Y. 2014. News credibility evaluation on microblog with a hierarchical propagation model. In *2014 IEEE International Conference on Data Mining (ICDM)*, 230–239. IEEE.
- [Jin et al. 2015] Jin, Z.; Cao, J.; Zhang, Y.; and Yongdong, Z. 2015. Mcg-ict at mediaeval 2015: Verifying multimedia use with a two-level classification model. In *Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop*.
- [Jin et al. 2016a] Jin, Z.; Cao, J.; Zhang, Y.; and Luo, J. 2016a. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*.
- [Jin et al. 2016b] Jin, Z.; Cao, J.; Zhang, Y.; Zhou, J.; and Tian, Q. 2016b. Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia* in publish.
- [Kwon et al. 2013] Kwon, S.; Cha, M.; Jung, K.; Chen, W.; and Wang, Y. 2013. Prominent features of rumor propagation in online social media. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, 1103–1108. IEEE.
- [Le and Mikolov 2014] Le, Q. V., and Mikolov, T. 2014. Distributed representations of sentences and documents. In *ICML*, volume 14, 1188–1196.
- [Mikolov et al. 2013] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.
- [Morris et al. 2012] Morris, M. R.; Counts, S.; Roseway, A.; Hoff, A.; and Schwarz, J. 2012. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 441–450. ACM.
- [Robertson and Zaragoza 2009] Robertson, S., and Zaragoza, H. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.
- [Salton, Fox, and Wu 1983] Salton, G.; Fox, E. A.; and Wu, H. 1983. Extended boolean information retrieval. *Communications of the ACM* 26(11):1022–1036.
- [Tumasjan et al. 2010] Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Weppe, I. M. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM* 10:178–185.
- [Wang et al. 2012] Wang, H.; Can, D.; Kazemzadeh, A.; Bar, E.; and Narayanan, S. 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, 115–120. Association for Computational Linguistics.
- [Wu, Yang, and Zhu 2015] Wu, K.; Yang, S.; and Zhu, K. Q. 2015. False rumors detection on sina weibo by propagation structures. In *IEEE International Conference on Data Engineering, ICDE*.
- [Zhao, Resnick, and Mei 2015] Zhao, Z.; Resnick, P.; and Mei, Q. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the*

24th International Conference on World Wide Web, 1395–1405. International World Wide Web Conferences Steering Committee.